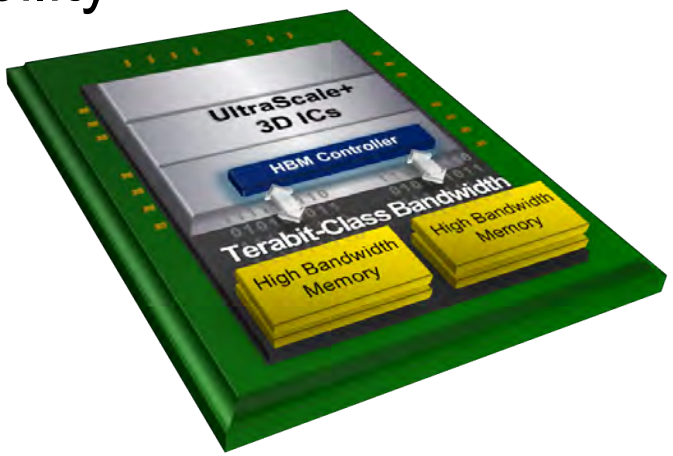
2.5D FPGA-HBM Integration Challenges

Jaspreet Gandhi, Boon Ang, Tom Lee, Henley Liu, Myongseob Kim, Ho Hyung Lee, Gamal Refai-Ahmed, Hong Shi, Suresh Ramalingam

Xilinx Inc.,
San Jose CA

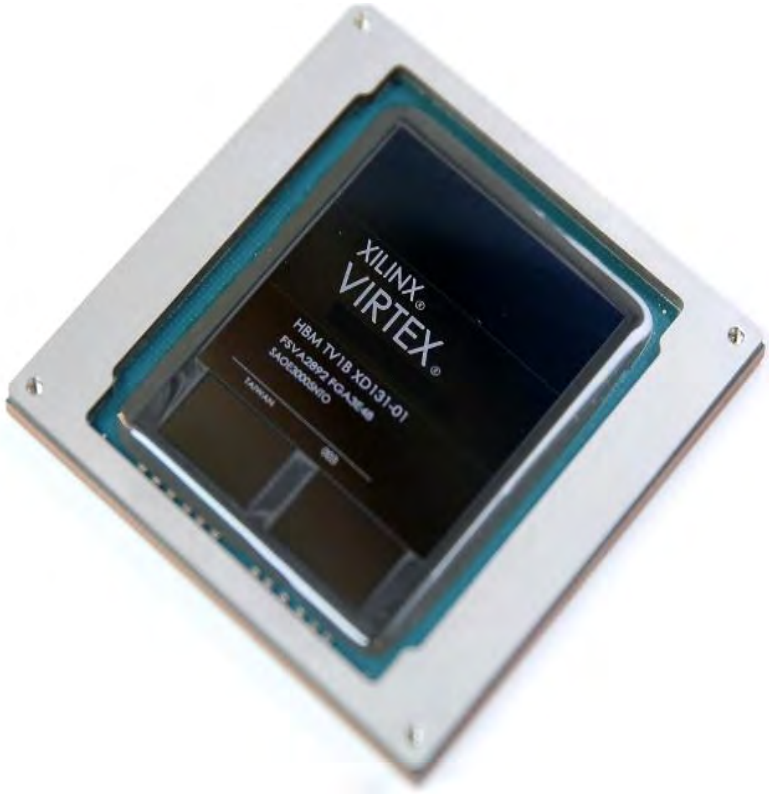
Presentation Outline

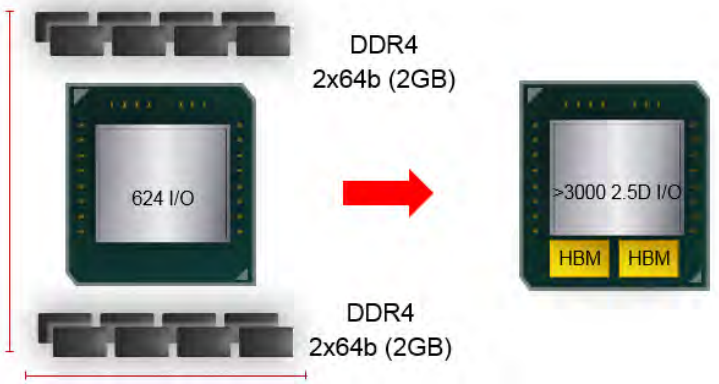
- **What/Why**
  - –Product Introduction & Motivation
- **How**
  - –2.5D Interposer Design & HBM Considerations
  - –CoWoS Process Integration & CPI
  - –Thermal Challenges
  - –SiP Component & Board Level Reliability
- **Summary**



Virtex 16nm UltraScale+ FPGA-HBM Product

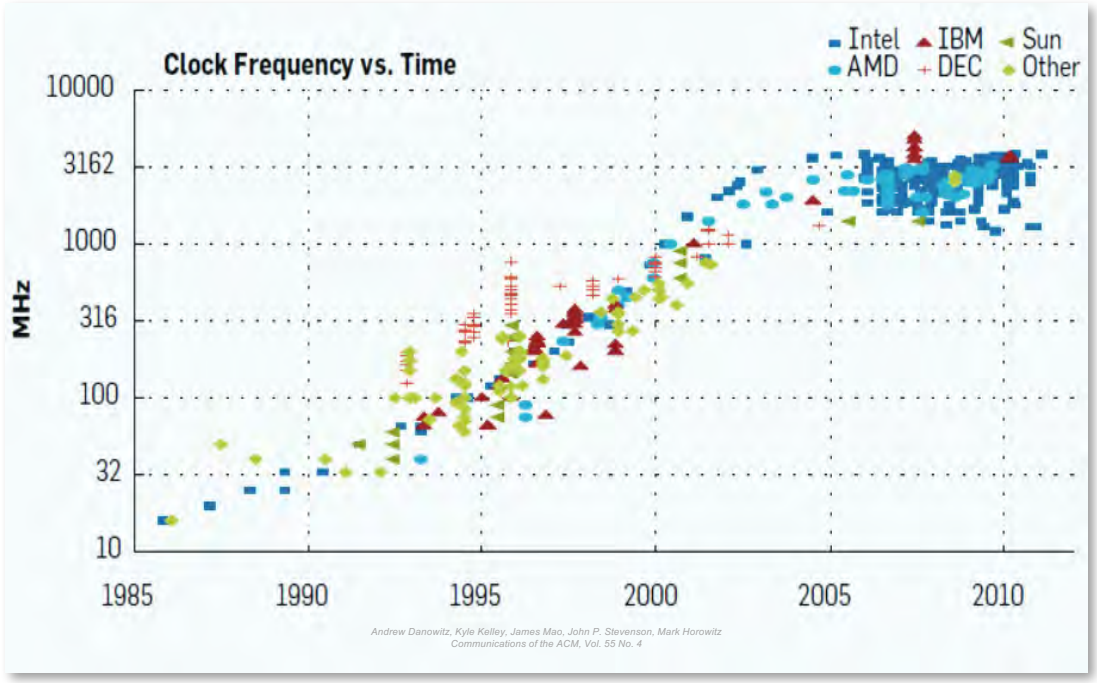
- Partitioned FPGA co-packaged with stacked DRAM (HBM) using Xilinx 3rd Gen Stacked Silicon Interconnect Technology (SSIT) based on CoWoS platform
- Revolutionary increase in memory performance delivering **10x bandwidth per HBM stack** and **4X lower power vs DDR4**
- Reduced board space and complexity
- 55mm^2 Lidless package for enhanced thermal performance, < 12mil coplanarity
- Copper Pillar C4 bump with Pb-free solder for fine pitch interconnect to substrate
- Passed JEDEC component & board level reliability





CPU Architectures not Scaling with Workloads

- Processor frequency scaling ended in 2007
- Multicore architecture scaling has flattened

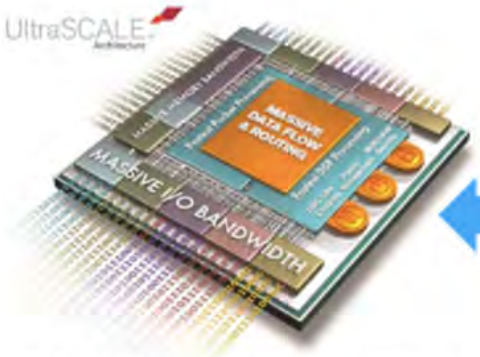


- Workloads require higher performance, lower latency
  - Cloud: video, big data, AI…
  - Edge: auto, surveillance, AI…

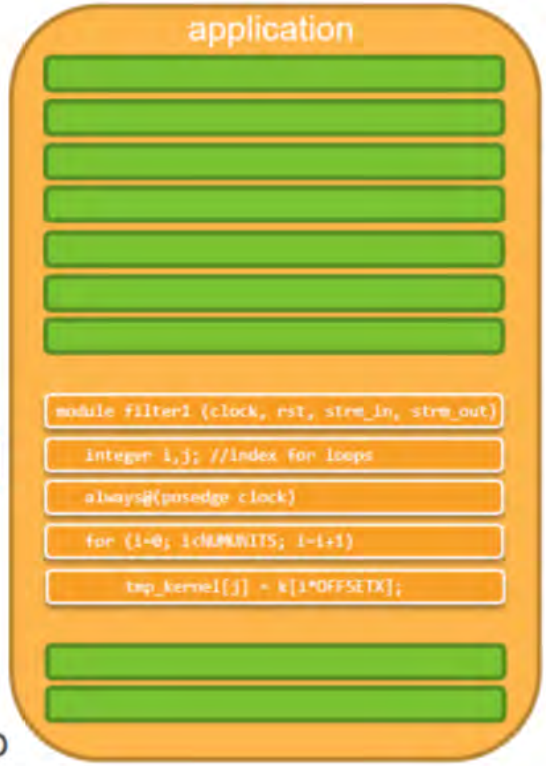
- Heterogeneous compute architectures needed
- Processors need to offload the compute intensive tasks to application specific accelerators that can provide performance and low latency

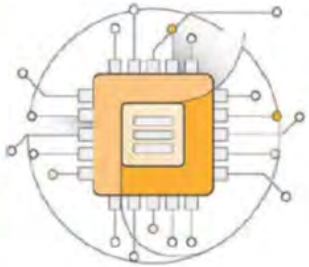
How FPGA Acceleration Works on AWS

FPGA handles compute-intensive, deeply pipelined, hardware-accelerated operations



Dedicated PCIe and ring connections also allow communication between up to 8 FPGAs, at up to 400Gbps





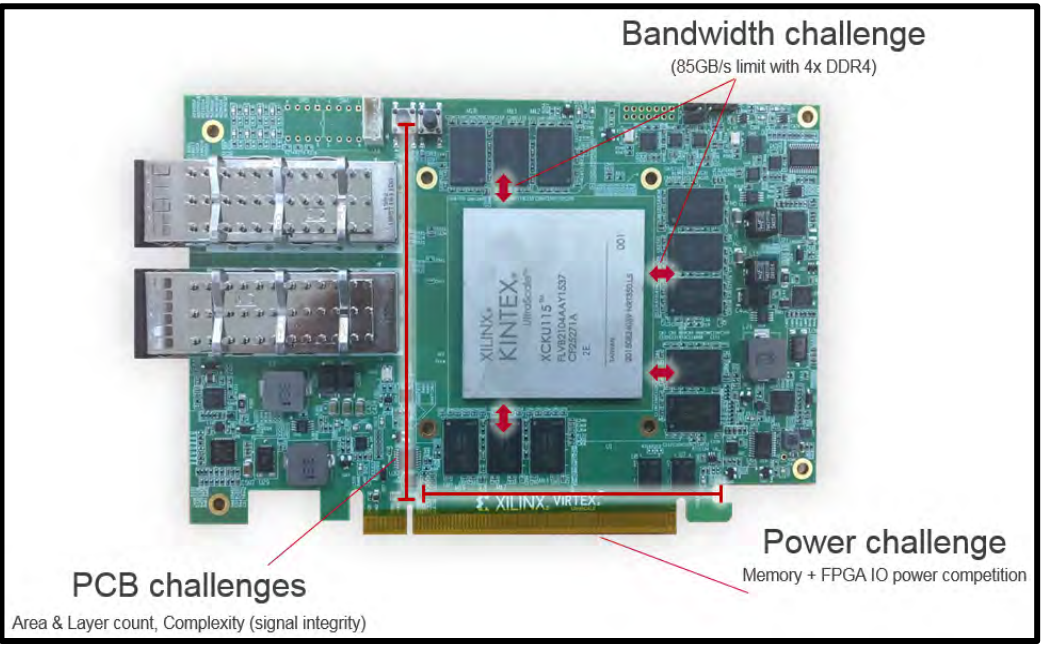
CPU handles the rest

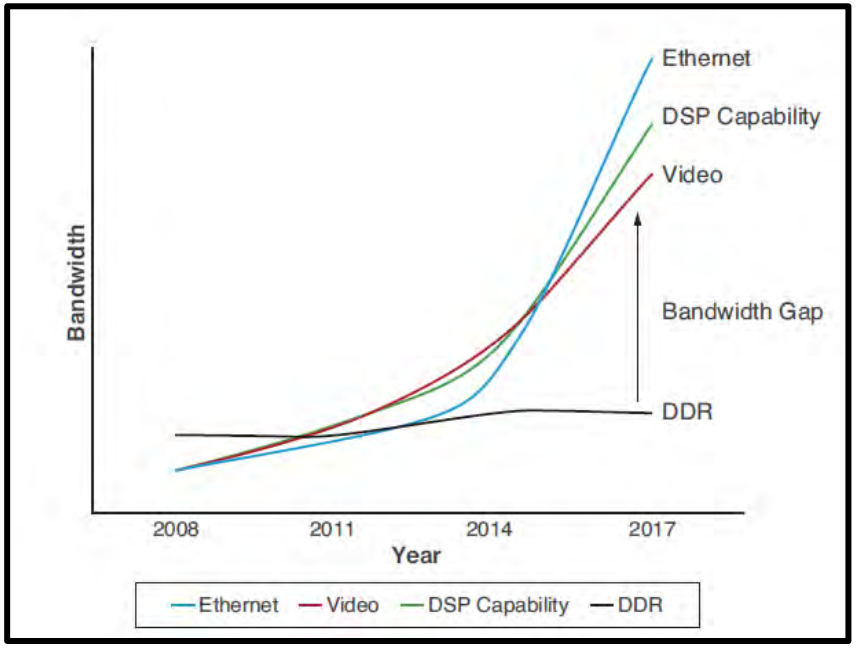
Data is transferred to and from the FPGA via PCIe

API's are run on the CPU to reprogram the FPGA to accelerate the workload as needed

Acceleration Requires Lot of Memory BW

- DDR4 data rate today **less than 2X** what DDR3 could provide in 2008





- Thanks to TSV die stacking, memory wall has been broken (for now)

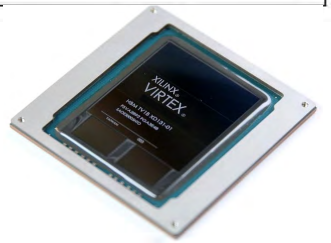
Memory Technologies Today

	DDR-4 DIMM	RLDRAM-3	HMC	HBM
Description	Standard commodity memory used in servers and PCs	Low latency DRAM for packet buffering applications	Hybrid memory cube serial DRAM	High bandwidth memory DRAM integrated into the FPGA package
Bandwidth	21.3GB/s	12.8GB/s	160GB/s	460GB/s
Typical Depth	16GB	2GB	4GB	8GB
Price / GB	$	$$	$$$	$$
PCB Req	High	High	Med	None
pJ / Bit	~27	~40	~30	~7
Latency	Medium	Low	High	Med

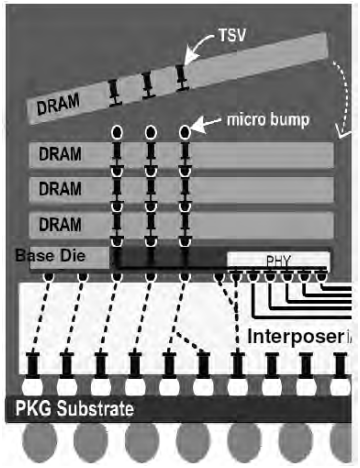
Everywhere

Wired Comms

Data Center



Data Center + Wired Comms

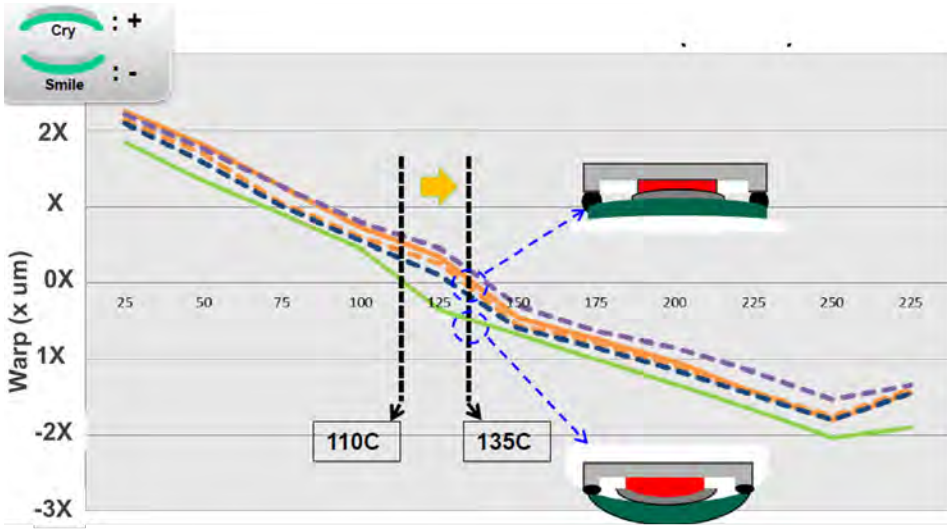


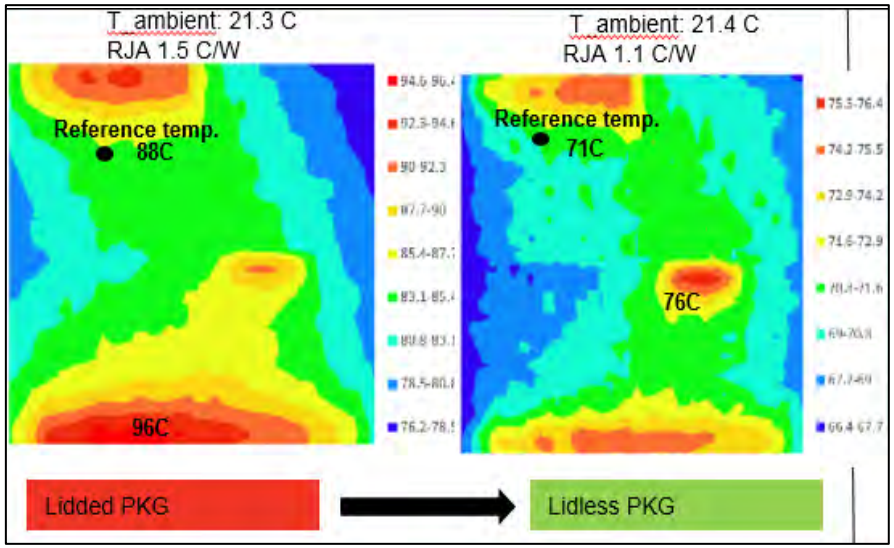
- **High Bandwidth Memory (HBM) is a new type of memory integration technology that vertically stacks memory chips via TSVs (thru silicon vias) providing low power consumption, ultra wide communication lanes, faster speed and smaller form factor**

Pic Source: http://cdn.wccftech.com/wp-content/uploads/2014/09/HBM.jpg

Why Lidless Package ?

- Programmable logic capacity growing 2-3X every 2-3 years
- But device/package size is not growing
- Increasing Power Density Driving Thermal Management Innovation





Thermal enhancement by moving to lidless pkg.

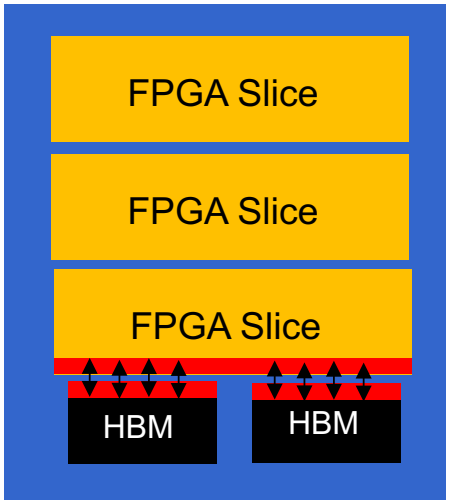
Thinner TIM
Poor Coverage

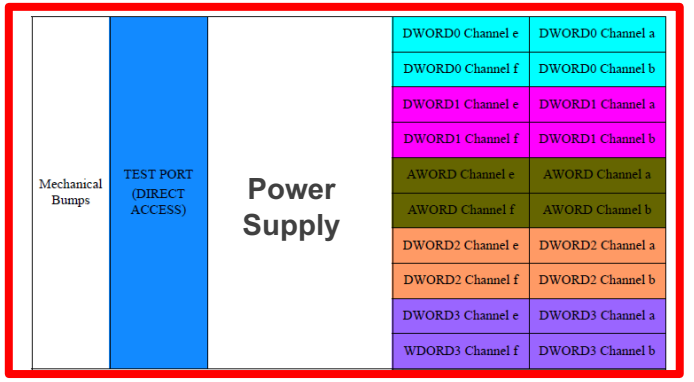
Good Coverage
Thicker TIM

How ?

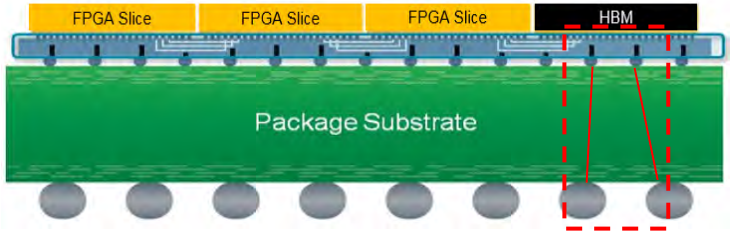
Interposer Design Considerations

- **FPGA PHY and HBM PHY ubump pitch must match for signal timing and uniform routing**
  - *Different mask design, Plating non-uniformity, D2I Bond line*
- **Open space between dies dictated by electrical signal integrity and CPI rules**
  - *Wafer & chip module warpage causing C4 opens/bridging, Underfill Flow dynamics*
- **Sufficient metal routing layers, minimal routing length & resistance, careful shielding of high speed signal lines required to minimize electrical cross-talk**
- **HBM cube comes with a set of direct access (DA) ports which have to be routed to BGA balls for RMA purpose**
  - *Routing Constraints, DA ports vendor specific*





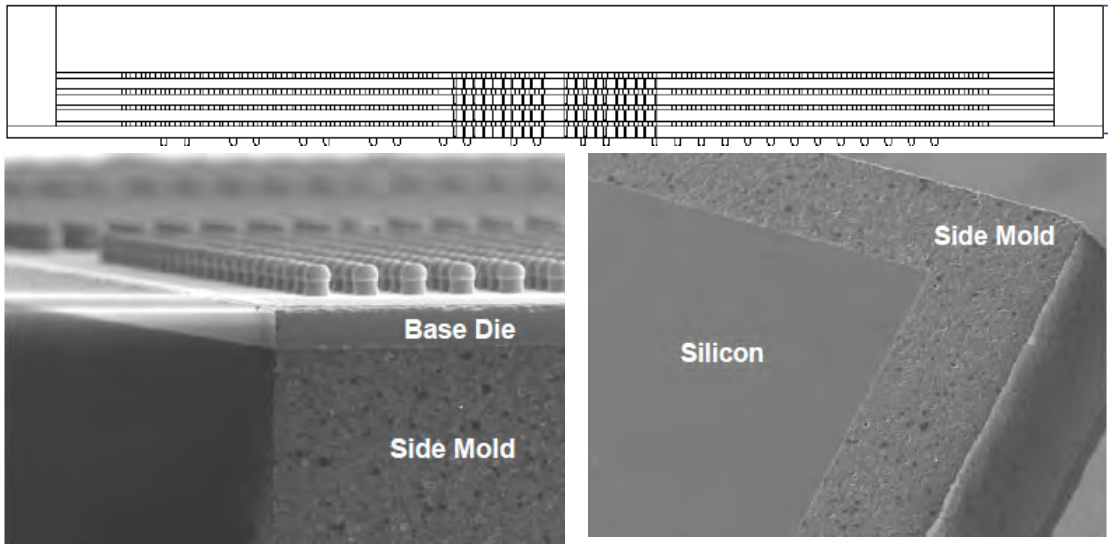
HBM buffer die layout (partial picture)



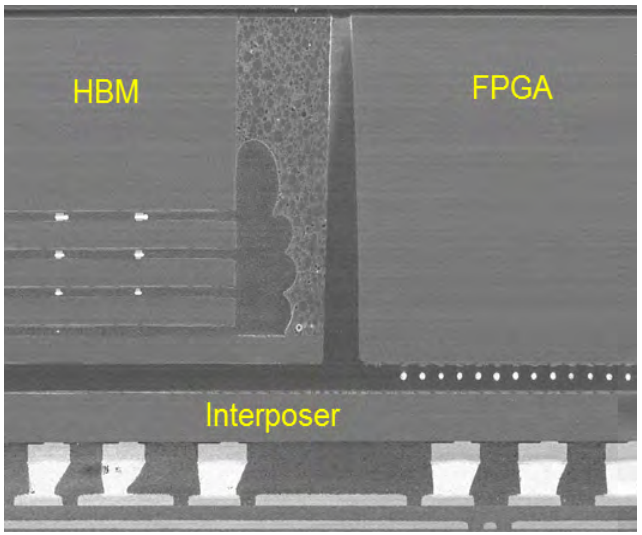
HBM DA Balls

HBM Vendor Selection & Swap ➔ Key Considerations

S. No	Considerations	JEDEC Std.	Impact
1	Package Fiducial	Yes	
2	Buffer die ubump layout/pitch/dimensions	Yes	
3	Package Size	No	SiP Design, Thermal, Warpage
4	Core die size	No	Warpage
5	ubump shape/metallurgy/coplanarity	No	Reliability, Yield
6	Vendor HBM Test Environment	No	SiP Electrical Design
7	DA port count/assignment/location	No	SiP Design, Test Board Design
8	Operation Temp. Range	No	Customer, Reliability
9	Memory Tech Node	No	Customer, Product Longevity

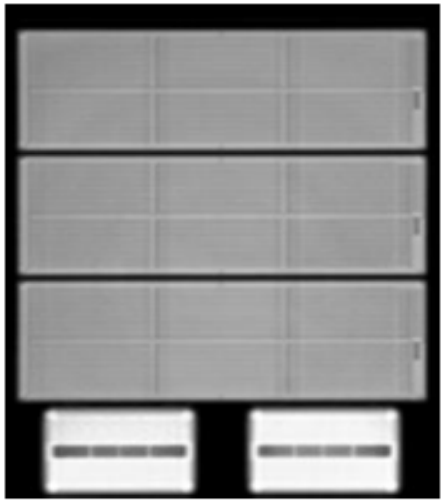


Images from Hynix presentation in Semicon Taiwan 2015

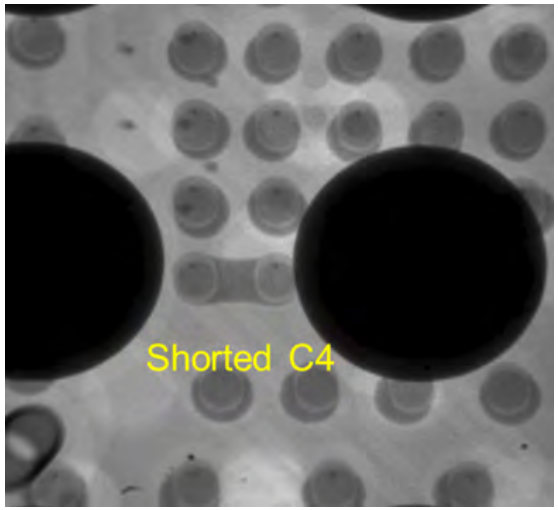


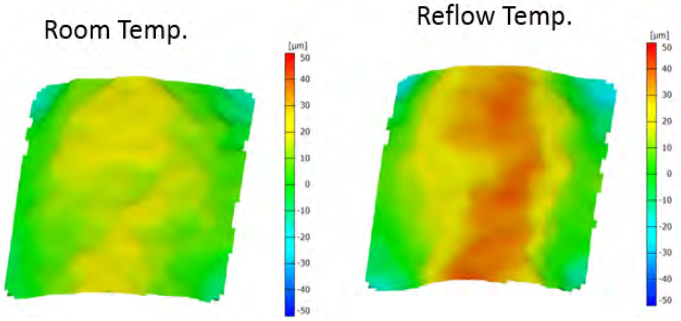
Xilinx TV

CoWoS Process Integration



- Xilinx 2.5D HBM-FPGA integration cover 2 corners of a super-large interposer (~1300mm2) with tighter C4 pitch
- **Concerns:** C4 opens/shorts due to high warpage caused by interposer open areas and asymmetric structure
- Different warpage behavior➔ FPGA-2 HBM CoW or CoC die has different warpage curvature than a SoC-4 HBM die
  - *C4 bump and substrate pre-solder size optimization*
  - *CoW die warpage reduction with underfill selection*





CoW die warpage at different temps.

ubump underfill	UF # 1	UF # 2
Die warpage at 250C, um	70	50

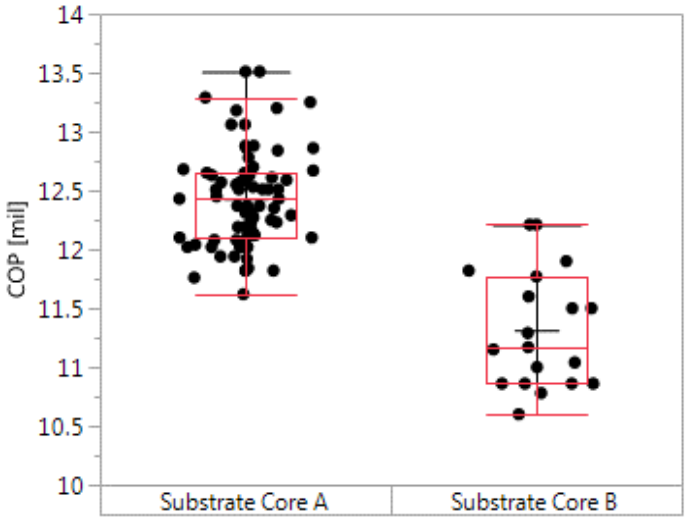
CPI Considerations & Mech. Design

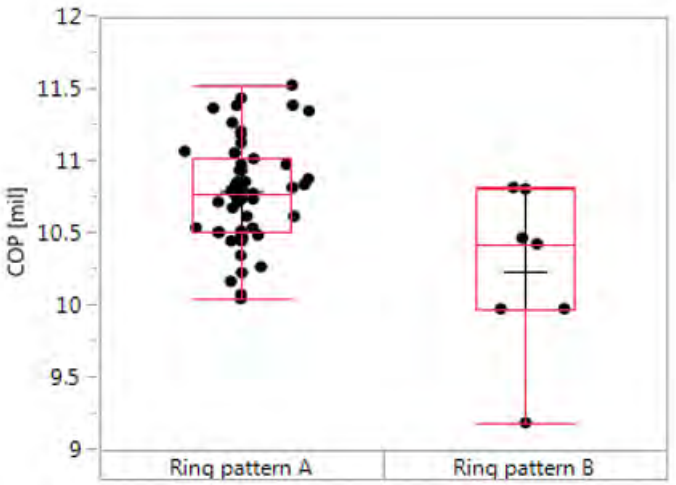
Copper Pillar Bump (CPB): Fine pitch interconnect, bump reliability, and pkg. thermal performance

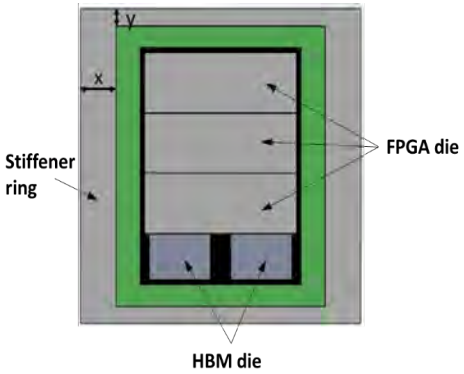
- **Concerns: Increased package stress due to high Tg underfill➔ Delamination, Cracking**
  - *Underfill material selection, curing, interposer dicing, etc. can help improve CPI performance*

Stiffener ring: Thermal performance & reduced cost

- ***Concerns: Combination of CPB & ring➔ Higher package coplanarity***
  - *Thicker & lower CTE substrate core material can help but BGA board level reliability impacted*
  - *Stiffener ring design, adequate adhesive material can help but heat sink assembly and KOZ between ring & chip capacitors impacted*



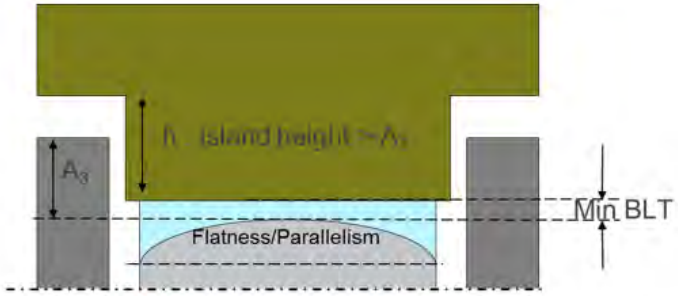


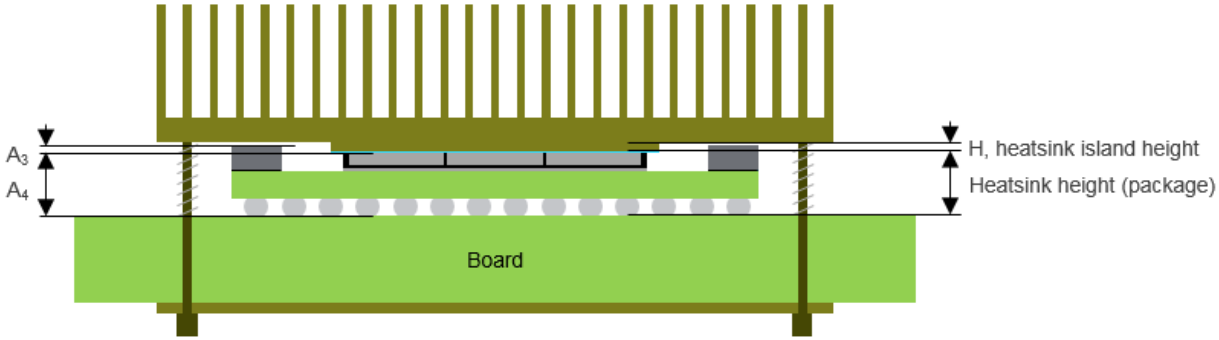


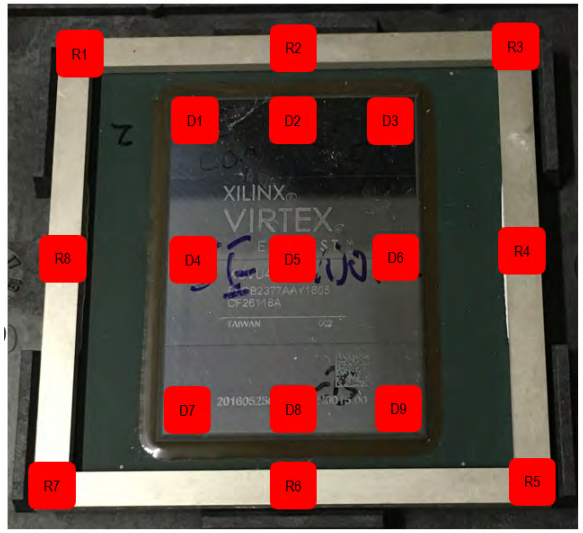
Ring thickness (Z, mm)	Ring thickness A- 0.2mm	Ring thickness A	Ring thickness A+ 0.2mm
COP (mil)	**12.4**	**11.5**	**11.1**
Ring width (X, mm)	Ring width A- 1mm	Ring width A	Ring width A+ 1mm
COP (mil)	**12.5**	**12.1**	**11.5**

New Process Metrics for Lidless Package

- **Current industrial practice**
  - *Lid tilt*
  - *Package coplanarity*
- **New metrics for stiffener ring**
  - *Flatness/Parallelism ➔ Enable lowest TIM BLT*
  - *Delta (A3) between Die & Stiffener ➔ Ensure no interference between heatsink/stiffener*





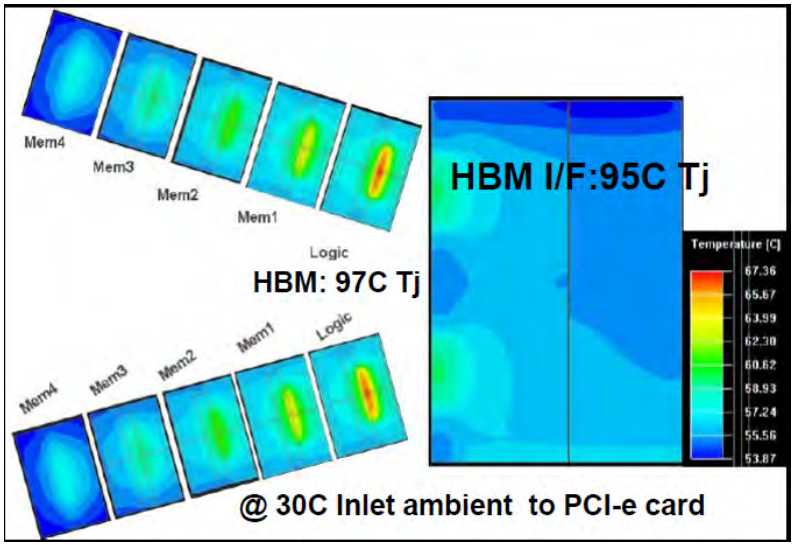


Measurement point

$$\textbf{Flatness} = \max(\textbf{D1}:\textbf{D9}) - \min(\textbf{D1}:\textbf{D9})$$
$$\textbf{Parallelism} = \max(\textbf{D2}, \textbf{D4}, \textbf{D5}, \textbf{D6}, \textbf{D8}) - \min(\textbf{D2}, \textbf{D4}, \textbf{D5}, \textbf{D6}, \textbf{D8})$$
$$A_3 = max(R1:R8) - min(D1:D9)$$

Thermal Challenges

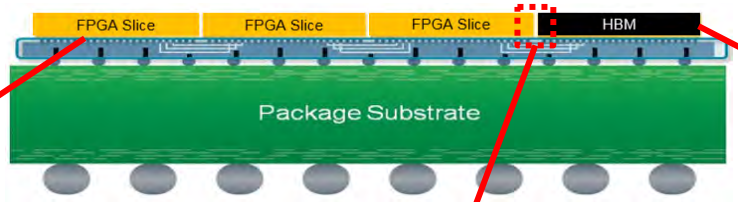
- **FPGA performance gated by HBM memory Tj limit: 95C (EM lifetime reduced at 105C)**
  - For 24/7 operation with T_a = 50C ➔ FPGA 100 C, Memory 103C
  - For 10% operation with T_a = 60C (AC failure)➔ FPGA 110 C, Memory 113C
  - HBM gradient ~10C (~2C/Layer), 8-Hi will be a challenge

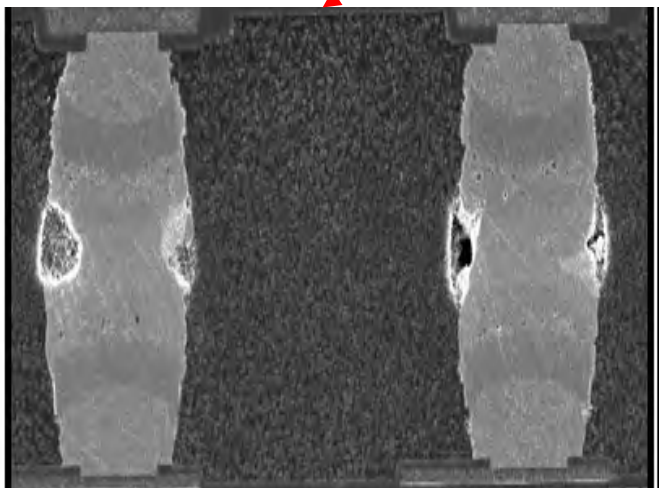


- **Close collaboration required**
  - Drive memory vendor for 105C operation
  - Highly conductive TIM
  - Co-work with customers for efficient cooling solutions

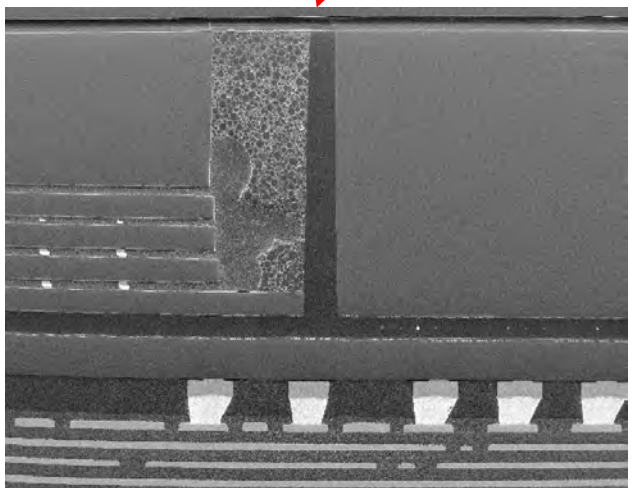
Pkg. Level Reliability

Test	Condition	Sample Size	Pre-con (MSL4)	96h	264h	432h	850X	1000X	1200X
HTS	150C	85	85/85	NA	NA	NA	NA	85/85	85/85
u-HAST	110C/85% RH	74	74/74	74/74	74/74	74/74	NA	NA	NA
TC-G	-40C to 125C	85	85/85	NA	NA	NA	85/85	85/85	85/85

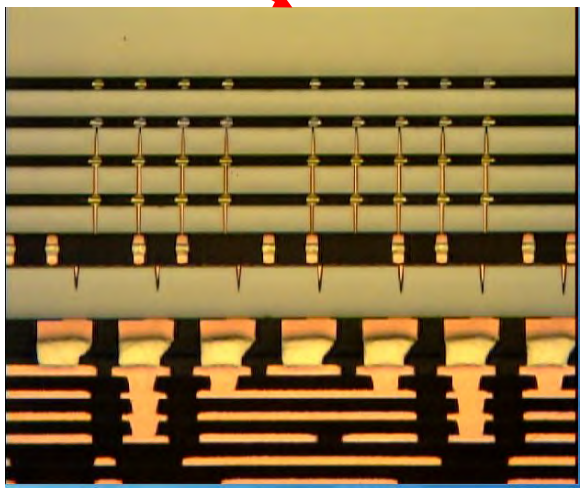




DMV ubump
HTS 1000 hrs



HBM - DMV gap
uHAST 264 hrs

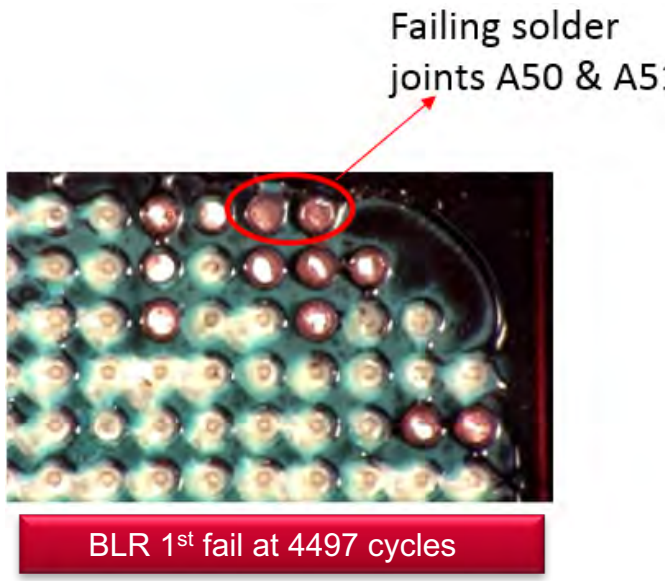


HBM on interposer
TC-B 1000X

Board Level Reliability

Bottom Material	BLR Schedule (cycles) (0 to 100C)				
	Cycles Completed	# Component Tested	# Failed	1st Failure	Char Life (cycle)
Meg 6	6000	16	1	4497	5476
New Material	6000	16	1	4883	5537

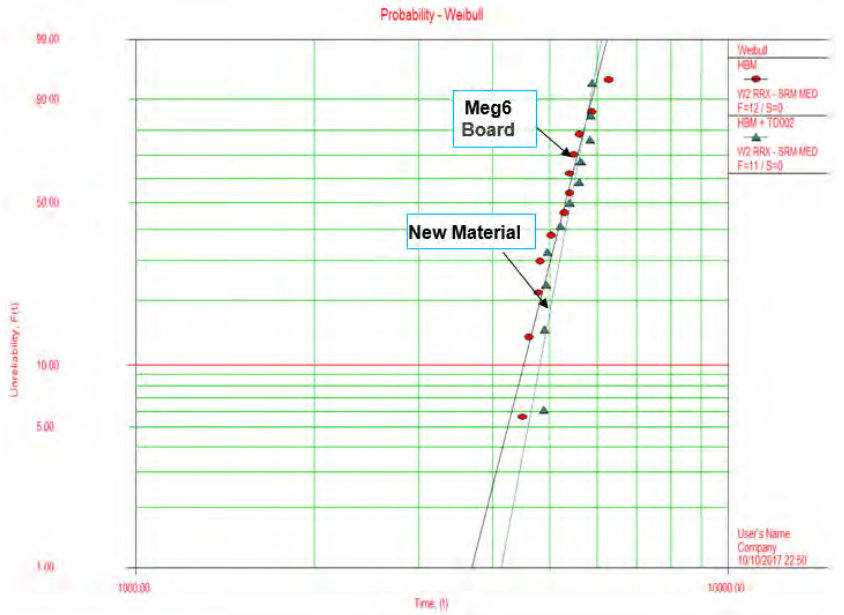
- **BLR test (0 to 100C):** Passed over 4000 cycles. Dye and Pry on the failed unit showed solder ball cracking at the package corner BGA balls. The solder cracks were on the package side
- **Shock test:** Passed both 100G (Cond. C) and 200G (Cond. D). Dye & Pry showed no solder cracks
- **Bend Test:** Complete with global strain ranging from 3639 to 4246 ue (micro-strain)



Board Level Reliability

Bottom Material	BLR Schedule (cycles) (0 to 100C)				
	Cycles Completed	# Component Tested	# Failed	1st Failure	Char Life (cycle)
Meg 6	6000	16	1	4497	5476
New Material	6000	16	1	4883	5537

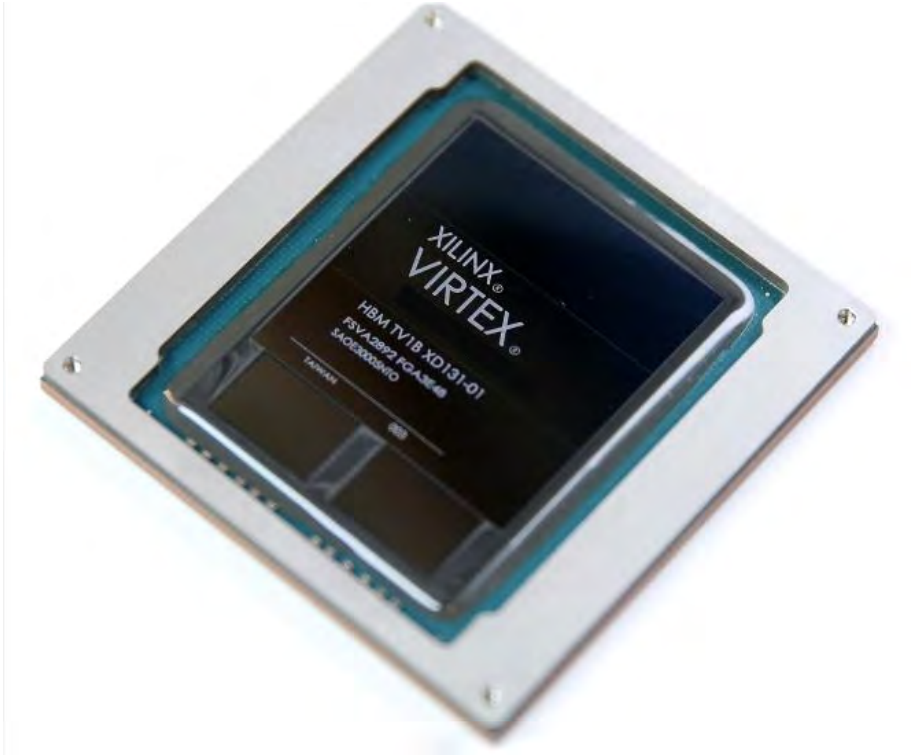
- **BLR test (0 to 100C):** Passed over 4000 cycles. Dye and Pry on the failed unit showed solder ball cracking at the package corner BGA balls. The solder cracks were on the package side
- **Shock test:** Passed both 100G (Cond. C) and 200G (Cond. D). Dye & Pry showed no solder cracks
- **Bend Test:** Complete with global strain ranging from 3639 to 4246 ue (micro-strain)



No significant difference between new & standard material

Summary

- **Low latency bandwidth and lower system power is driving the need for die partition and HBM adoption**
- **Heterogeneous SiP design & performance gated by HBM constraints**
  - DFx approach & close knit collaboration required between memory vendor, design, process, test and external customers
- **To drive broader adoption of HBM applications (cooling limited) and higher performance stacks (8-Hi), higher HBM junction temperature (>95C) needs to be supported**
- **Package substrate material selection & stiffener ring design are key enablers to meet component coplanarity, reduce thermal resistance and achieve high reliability for a large body lidless package**



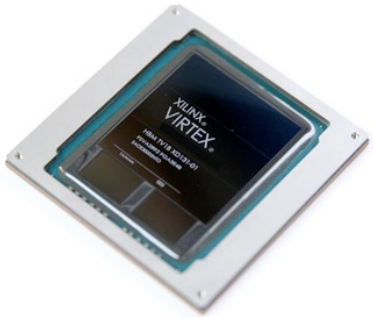
Thank You !

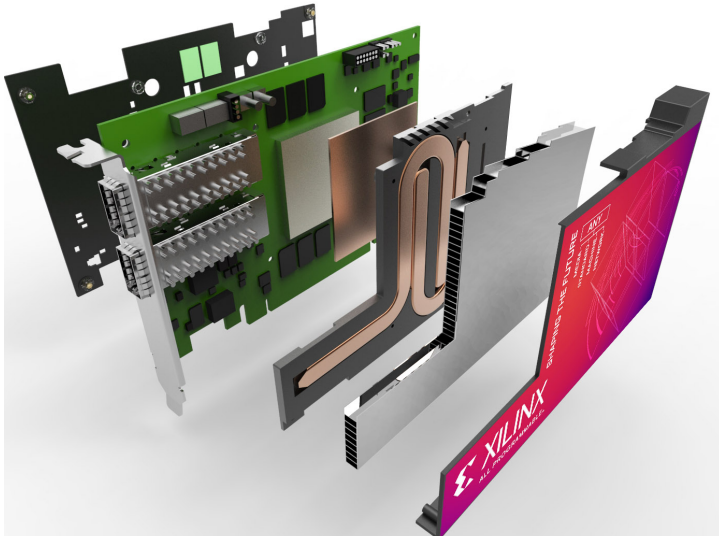
Appendix

Not Discussed

- FPGA & HBM Vendor Rules of Engagement
- HBM IQC
- SI, PI, Timing Challenges
- Test Hardware Challenges
- Electrical Test Data
- Thermal Details

FPGA-HBM Target Applications





Wired
(200G – 800G)

T&M
(Testers, AWG)

A&D
(Digital RF Memory)

AVB
(8K Video)

Compelling Use-Cases for Acceleration

Deep Learning Training and Inference

Video and Image Processing

Engineering Simulations

Financial Computing

Molecular Dynamics

VR Content Rendering

Accelerated Search and Databases

Many More

Ever Increasing Power Density

- SoC Are Growing, Fast
  - Programmable logic capacity growing 2-3X every 2-3 years
  - Heavy Hard-IP (SoC) content driving up power density
  - "More than Moore" 2.5 and 3D IC Technology
  - But device/package size is not growing
    - More than doubling the capability in the same footprint
    - More Integration in Device (Logic, memory, Optical, VR…)
- System Level (PCI-e, Server)
  - Fixed power
  - Fixed form factor
  - Same environment
- Increasing Power Density Driving Thermal Management Innovation
  - This is why Xilinx is very focused on improving thermal design